Journal of Climate Change, Vol. 8, No. 3 (2022), pp. 51-62.
DOI 10.3233/JCC220021

Performance Analysis of Ensemble Techniques for Rainfall Prediction: A Study Based on the Current Atmospheric Parameters

Geeta Mahadeo Ambildhuke* and Barnali Gupta Banik
Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation
Deemed to be University of Hyderabad, Telangana – 500075, India
✉ geeta@klh.edu.in

Received July 12, 2022; revised and accepted August 2, 2022

Abstract: Rainfall prediction is the most significant requirement nowadays due to the chaotic nature of climate. Climate has changed drastically over the last few years due to global warming and has become very unpredictable. Rainfall prediction is essential for decision-making in various sectors like agriculture, transportation, tours and travels, outdoor events, etc. In this study, machine learning algorithms are analysed and experimented on the dataset comprising various atmospheric parameters of Hyderabad city in Telangana. The work is carried out on individual popular classifiers, namely Naïve Byes, Decision Tree, Random Forest, K nearest neighbour, and Support Vector Machine. The performance is compared with techniques like voting classifiers and stacking ensemble. The experiment gives predictions on the rainfall intensity as No Rain, Low to Medium rain, or Heavy rain. The k-cross-fold validation technique is used as the evaluation metric, which is very effective and results in less biased estimations. The aim is to provide the decision-making capabilities based on the mentioned intensity of rainfall that can be very useful in managing the irrigation cycle in the agriculture field, deciding on an outdoor event, or any travelling plan based on the current atmospheric parameters. The platform used is python, which is portable, open to access, and available easily.

Keywords: Rainfall prediction; Machine learning algorithms; Ensemble techniques; Nowcast; Weather forecasting; Climatic parameters.

Introduction

Water is the basic necessity of every living thing and must be utilised most appropriately. It is one of the scarce resources as well. As we all know, the agriculture sector uses 70% of the available water. Precision Agriculture (PA) is the new emerging agriculture practice that manages the field and primary resources like water, pesticides, and fertilisers, precisely, at the right time, at the right place, and in the right amount. To make such a decision, weather prediction plays a significant role in agriculture; if the rain status is known beforehand, the irrigation schedule can be managed, and the farmer can decide whether to irrigate the field or not or to what extent, depending upon the intensity of rainfall.

Due to global population expansion and rising world economies, the need for world agriculture output will grow enormously in the following decades. Simultaneously, changes in demand and climate will impact the agriculture sector, putting the world's agriculture resources' productive capability to the

*Corresponding Author

test. Hence, development in the agricultural sector is required to improve its ability to supply this growing global demand.

Technological advancements like the Internet of Things (IoT), wireless Sensor networks (WSN), cloud computing, and meteorological satellite helped collect lots of weather data available for weather forecasting In particular, big data analytics play a significant role in predicting the climate based on data from the past. As the data is collected from various sources, it is available in an unstructured manner that contains lots of information. To make proper use of such data, techniques like data mining are very important to extract the appropriate information from such sources. In addition to this, technologies like machine learning (ML) and artificial neural (AI) networks can transform such information into a useful format and presentable way for needy people as per their domain to enhance their decision-making capability for the growth or to take preventive measures against any unappropriate situations. An accurate rainfall prediction has become more crucial due to climate variations (Cramer et al., 2017). Machine learning methods and Deep learning techniques could forecast the rainfall by extracting hidden patterns from historical data among weather attributes.

Climate prediction is still in its initial stages. Much research and development are going on for a reliable climate prediction as the weather is changing its patterns and has become chaotic in the last few years due to global warming, which is the effect of pollution that has increased the emission of carbon dioxide in a large amount. Deforestation and using chemicals are some of the main reasons that make climate change its pattern. Many unpredicted events have been experienced that has led to the loss of lives and property.

Crops need water in their different stages of growth, from sowing to harvesting, therefore, the agriculture sector is extremely dependent on the daily weather for the status of rainfall. The weather predictions support farmers in decision-making for planning the irrigation based on the status of rainfall to optimise the yield. Irrigation scheduling for a crop can be managed by predicting the rainfall based on weather parameters at that location. This will help determine the necessary amount of water to be supplied to the crop. Rainfall based on actual parameters at that location will be helpful, thereby avoiding excess irrigation and water stress if no rainfall occurs. However, predicting rainfall is still a problematic endeavour. Therefore, it is crucial to use appropriate methods for categorising rainfall over a region. In the meantime, ML techniques have been suggested to improve the precision of rainfall predictions (Meyer et al., 2018).

Most experiments were conducted to predict and categorise the rainfall as rain or no rain; including this, some experiments classify the day as sunny, cloudy, or rainy. Atmospheric data used in this study is for Hyderabad city situated in Telangana, India. Data set is collected from the Nasa power. Data was accessed as daily data for Hyderabad city located at Latitude: 17.3954 and Longitude: 78.4504 for the period of 20 years from the year 1981 to May 2021. Data pre-processing is done by analysing the data collected. Manually the data is balanced and labelled based on the amount of precipitation as no rain to deficient rain, low to medium rain, and medium to high rain.

Many classification algorithms such as Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbour (KNN), and others have been investigated for the prediction of rainfall and compared with ensemble techniques based on the time taken and accuracy. Because of how differently these algorithms function from one another, there is a possibility for more improvement by adjusting training and testing ratios or combining other strategies.

The main objectives are:

- To determine the rainfall based on the intensity, like no rain, low to medium rain, or high rain.
- Most popular classifiers are experimented with and compared with the various ensembled techniques and explored the best of them.
- This rainfall prediction model will benefit the people whose work area is mostly dependent on the rain, like farmers, travellers, transporters, outdoor event management systems, etc.

This study is further divided into the following sections. The Related Surveys section provides information about the most recent work in the related field of weather prediction. Section on Existing Knowledge of Weather Forecasting highlights the knowledge of weather forecasting, its meaning, the difference between weather and climate, and why weather forecasting is needed. Section on Collection of Data and Data Pre-processing gives an insight into the dataset, data collection, and various steps required in pre-processing data. Methods and Methodologies section describes various methods and methodologies used to experiment. The Results and Analysis section shows the results, their analysis, and comparison to

know the best among those, and finally, Conclusion and Future Scope section ends with the conclusions and prospects.

Related Surveys

This section presents the work done by some researchers in the field of weather predictions and will present different methods and methodologies.

Manandhar et al. (2019) included seasonal, diurnal, and ground-based weather parameters. They identified the most required parameters for predicting rainfall and gave them as input to machine learning algorithms. The surface weather parameters like humidity, temperature, dew point, etc., are considered along with precipitable water vapour (PWV) and is defined as a measure of the total water vapour stored in a column of the atmosphere obtained from the GPS and worked on various features by eliminating and adding features and concluded that the combination of various parameters helps to reduce the false alarm rate and increase accuracy.

Misra et al. (2018) presented a significant goal to find suitable concurrent meteorological characteristics concerning pressure depth in the atmosphere and then use these meteorological parameters to simulate daily rainfall. A feedforward multilayer perceptron (MLP) model is suggested for this purpose. MLP model is used to simulate daily rainfall over India's central monsoon region in this study. Four concurrent meteorological parameters, namely geopotential height, specific humidity, zonal wind, and meridional wind, depend on daily rainfall. MATLAB platform is used to experiment on MLP. The model has shown limited performance in capturing the extreme condition of heavy rainfall and very low rainfall events.

Tharun et al. (2018) worked on a rainfall forecasting model for Coonoor, Nilgiris, and Tamilnadu. This work compares the performances of statistical modeling and regression techniques such as SVR, RF, and DT in terms of accuracy. From the observation, it is concluded that when compared to regression techniques, statistical modeling fails to provide good Accuracy for rainfall prediction due to the dynamic nature of the atmospheric composition.

Nastos et al. (2013) did work using three different ANN models to forecast rain intensity for the next four months. Dataset is obtained from the National Observatory of Athens (NOA). The model with ANNs shows good prediction but limitations in predicting peak rainfall intensity.

Gutierrez-Lopez et al. (2019) collected data from a historical database of 1237 storms, such as humidity, surface temperature, atmospheric pressure, and dewpoint, and a simple precipitation forecast model was suggested. The proposed model can use the proper combination of these characteristics to accurately forecast the time of storm onset. The findings show that the suggested methodology is effective.

Appiah-Badu et al. (2021) contributes to using several categorisation algorithms for rainfall prediction in Ghana's distinct ecological zones. Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), Extreme Gradient Boosting (XGB), and K-Nearest Neighbor (KNN) are some of the categorization techniques used. The Ghana Meteorological Agency provided the dataset, which included different climate parameters and covered 1980 to 2019. With different training and testing data ratios, the performance of the classification algorithms was evaluated based on execution time, f1-score, precision, recall, and accuracy. RF, XGB, and MLP fared well on all experimented training and testing ratios—70:30, 80:20, and 90:10—while KNN did the worst across all zones. DT is consistently portrayed as having the quickest model execution times, while MLP uses the most run time.

The current study by Pereira et al. (2022) seeks to evaluate climatic changes in the Ramanathapuram coastal region using the average monthly rainfall and temperature along the coastal region. Thirty years, or 1990 to 2019, have been used to analyse the microscale rainfall and temperature trend. According to the study, the region's 467 km^2 averaged 676 mm of rain over decade-I. About 39 km^2 of the study region received 637.6 mm of rain in decade II, 48 km^2 of the study area received 821 mm of rain, and only 29 km^2 received the typical amount of 992 mm of rain, while 351 km^2 had "excess" rainfall of over that amount.

Existing Knowledge of Weather Forecasting

In today's era, PA is the most promising approach toward farm management that makes use of technology (IoT) to observe the spatial and temporal parameters of the field and makes sure to use the input resources like water, fertilisers, pesticides, etc. to the crops at the right time, right place and in the right amount. The rainfall prediction based on weather parameters manages irrigation scheduling to determine the amount of water to be supplied through irrigation to maintain the threshold value of water (moisture) required by the crop at any time.

Climate and weather are generally described through many meteorological parameters like temperature, humidity, surface pressure, dew point, etc. Still, rainfall intensity or precipitation amount is the most important meteorological parameter to describe the climate or weather at any time (Abdullah et al., 2019).

The final yield of crops is determined by the environment in which they are grown. The uncontrollable environmental factors, such as climate and weather, have the most significant impact on crop productivity. When predicting the weather, scientists use scientific knowledge and technology to make weather observations and estimate what the atmosphere will be like in a specific location. It is a technique for anticipating events like cloud cover, rain, snow, wind speed, and temperature (Cahir, 2013).

Weather forecasting has different approaches based on various factors based on the period. It can be divided into four types: very short-term (real-time), short-term, medium-term, and long-term. Short-term and very-short term forecasts are more accurate than medium or long ranges. The real-time forecast ranges from 0-2 hours. In contrast, the period for short-term forecasting is 2 to 72 hours. "middle-term forecasting" refers to forecasting that extends beyond 72 hours and up to 240 hours. Long-term forecasting includes periods of more than 10 days, up to 30 days, and up to 2 years (Didal et al., 2017).

Furthermore, depending on the methodology used for predicting, forecasting systems can be categorised into two types: deterministic and probabilistic. Deterministic approaches provide exact weather forecast numbers for a specific location, whereas probabilistic methods suggest weather event probabilities. Again, deterministic forecasting can be further categorised depending on the models used as artificial intelligence or hybrid models (Jaseena et al., 2020). Based on meteorological parameters, the forecasting model can be divided into wind speed, temperature, precipitation, etc., as shown in Figure 1.

The Basic Difference Between Weather and Climate

Weather and climate are commonly confused, although they are not the same thing, despite having certain similarities. Temperature change, wind speed and

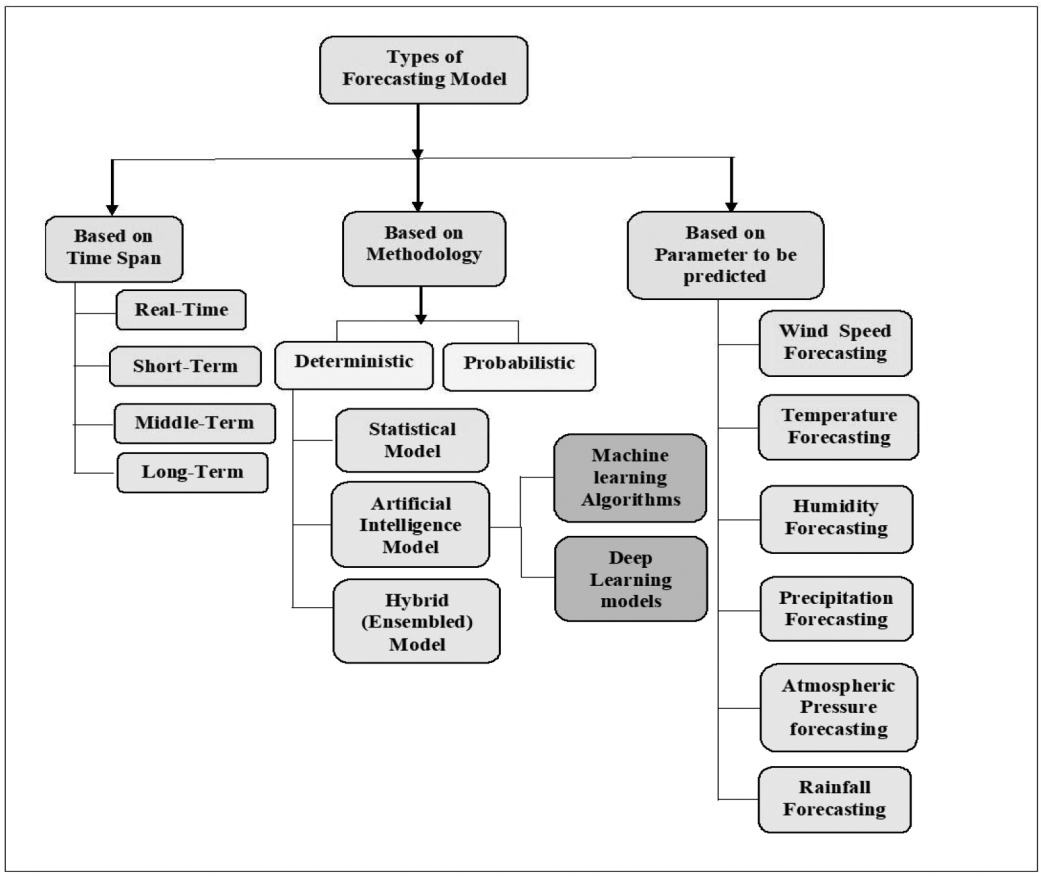

Figure 1: Classification of weather forecasting.

direction, levels of humidity, rain type and precipitation amounts, air pressure, types of clouds, and cloud coverage are all examples of weather and climatic elements. Weather is the term used to describe the everyday variations in the atmosphere or the state of the atmosphere over a short period. In contrast, when these weather patterns are combined, they form a climate for a particular location as averaged over many years. As per the observations, both climate and daily weather are changing their patterns due to the interference of humans and nature, which gives rise to the greenhouse effect and global warming, resulting in its chaotic nature.

Atmospheric Parameters Responsible for Rainfall

Atmospheric parameters such as temperature, humidity, wind, and pressure change abruptly, leading to instability in the atmosphere and may bring rain, develop storms, lightning, and thunder.

Temperature and humidity are the most important factors in predicting precipitation (Holley et al., 2014). However, four other meteorological variables are significantly linked to rainfall: air pressure, relative humidity (or dewpoint temperature), wind speed, and cloud cover (Lekouch et al., 2012).

Importance of Weather Prediction in Agriculture

Due to their timing, climate predictions of smaller weather events significantly affecting crop yields are also important. Most crops have developmental stages that are highly susceptible to weather conditions. Pollination for corn and pod filling for soybeans are two examples. Predicting crop yields and grain production levels can be improved by predicting climate conditions during these important development periods. Local climate forecasts increase agricultural production planning and, as a result, individual farmer yields (Agovino et al., 2019). Climate conditions on a global and regional scale influence farmers' agricultural prices and crop selection decisions.

Prediction of climate can be beneficial in creating a more precise fertiliser application plan for the growing season based on the values of temperatures and precipitation known during various phases of the growing season. Conditions like precipitation and temperature greatly influenced crops, pests, weeds, molds, insects, etc., in its growing season. So the negative impact of climate change on crop yield can be controlled or minimised by climate prediction in advance during various phases of crop growth. Similarly, humidity levels also affect the crop, and plants under stress are more susceptible to various living or non-living stresses. Corn yield can be significantly affected during the grain-filling phase if drought-like conditions are encountered.

Different Methods Adopted for Weather Forecasting

Modern weather forecasting is both exceedingly sophisticated and highly quantitative. The following are the numerous techniques used to predict the weather:

- Synoptic weather forecasting.
- Statistical approaches
- Numerical methods

Synoptic Weather Forecasting

Synoptic refers to a specific time of observation when describing the observation of several weather components. A weather map showing atmospheric conditions at a particular time is called a synoptic chart by meteorologists. This traditional way of forecasting the weather was used mainly during the 1950s. A meteorological center creates several synoptic charts daily to provide an average perspective of the shifting weather pattern. The fundamental building block of weather forecasts is synoptic charts. As previously said, creating synoptic maps regularly entails gathering and analysing a sizable amount of observational data from thousands of weather stations.

Some empirical guidelines were developed after years of careful examination of weather charts. These guidelines assisted the forecaster in determining the speed and direction of weather system migration. Synoptic approaches need an in-depth evaluation of recent weather reports from a wide geographic area. Forecasts are based on the premise that the current weather situation will act similarly to previous identical situations in light of the relationship between current weather patterns and those in the past. Choosing similar historical instances frequently relies on the forecaster's experience and memory, but with the arrival of computers, the process has become faster and more objective. Short-range forecasts benefit from using this methodology.

Statistical Approach

Statistical techniques rely on historical weather records under the presumption that the weather will be similar in the future. Finding meteorological features that are reliable predictors of future events is the fundamental goal of examining historical weather data. After

establishing these connections, accurate data can be used to anticipate future situations without risk. The information obtained by remote sensing satellites is typically used in weather forecasting at the macro level. Using photos captured by these meteorological satellites, weather characteristics such as maximum temperature, minimum temperature, the quantity of rainfall, cloud cover, wind speed, and their orientations are forecasted to determine future trends.

Time series analysis and future forecasting typically use the data mining technique known as autoregressive integrated moving average (ARIMA), a class of statistical models for assessing and predicting time series data. Forecasting climate change is crucial for protecting the planet against unforeseen natural disasters such as floods, frost, forest fires, and droughts. A class of statistical models for assessing and predicting time series data is known as an ARIMA model. It explicitly addresses several common time series data types and, as a result, offers a straightforward but effective technique for producing accurate time series forecasts.

Numerical Methods

Numerical Weather Prediction (NWP) is defined by Linacre and Geerts (2002) as a streamlined system of equations known as the fundamental equation used to calculate changes in circumstances. Numerical weather prediction is significantly used in contemporary weather forecasting. Meteorologists use enormous supercomputers with software forecasting models to generate weather predictions based on various atmospheric parameters, including temperatures, wind speed, high- and low-pressure systems, precipitation, snowfall, and other factors—the weather person analyses the data to make the weather forecast for the day. The accuracy of the forecast depends on the methods that the computer's software uses to forecast the weather. Errors result from equations that lack precision. Numerical weather prediction offers the most significant way to predict impending meteorological conditions compared to other methods.

Collection of Data and Data Pre-processing

The work is carried out for Hyderabad city, Telangana, and the dataset is collected from Nasa Power Data viewer for Daily data which contains many climate parameters. Still, the most relevant parameters are chosen, and the data is collected for 40 years from 1981 to May 2021. To collect enough data, a long time is to be taken to get sufficient data as there is an average of 64-65 rainy days in a year.

Rainfall in Hyderabad During a Year

Table 1 gives the average values of some important ground surface parameters for over a year for Hyderabad.

From Table 1, we can see those rainy days are on average 64, so while collecting the data, the no of rainy days is significantly less, and this data is divided into three categories as No_Rain to very Low Rain, Low to Medium Rain, and Medium to High rain. Figure 2 shows the rainfall pattern, which describes the amount

Table 1: Average atmospheric parameters month-wise for Hyderabad

Month	*Average annual rainfall (in mm)*	*No. of rainy days*	*Maximum temperature*	*Minimum temperature*	*Relative humidity (in %)*
JAN	13.2	0.7	28.8	15.2	56
FEB	7.9	0.8	31.9	17.6	49
MAR	15.3	1.0	35.4	20.8	39
APR	20.2	2.1	37.9	24.3	37
MAY	35.7	3.4	39	26.2	39
JUNE	103.8	10.0	34.5	24	61
JULY	169.9	12.4	30.8	22.6	71
AUG	178.7	14.1	29.8	22.1	74
SEPT	158.3	9.7	30.5	22.0	72
OCT	97.2	6.3	30.6	20.3	63
NOV	22.4	2.9	29.0	16.9	58
DEC	5.9	0.7	28.0	14.5	57
TOTAL		64.1			

of rainfall and number of rainy days in Hyderabad during the year.

Among various attributes (columns) in the dataset, the selected parameters (attributes) such as Max_ temperature, Min_Temperature, Temperature, Relative Humidity, Pressure, and Wind Speed are used as the independent variables, and the condition is used as the dependent variable and is labeled in three categories as no_rain to low rain, low to medium rain and medium to high rain based on the precipitation (precp) attribute in the dataset as shown in Figure 3.

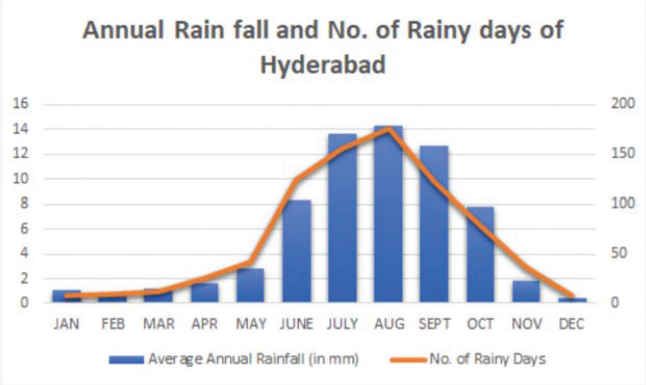

Figure 2: Amount of rainfall and number of rainy days in a year.

Dataset Division Based on Rainfall Intensity

The rainfall intensity can be classified based on the precipitation rate, which depends on the considered time. The following categories are used to classify rainfall intensity (Narvekar et al., 2015):

- Low rain — when the rate of precipitation is < 2.5 mm (0.098 in) per hour
- Medium rain — when the rate of precipitation is between 2.5 mm (0.098 in) and 7.6 mm (0.30 in) or 10 mm (0.39 in) per hour
- High rain — when the rate of precipitation is > 7.6 mm (0.30 in) per hour, or between 10 mm (0.39 in) and 50 mm (2.0 in) per hour

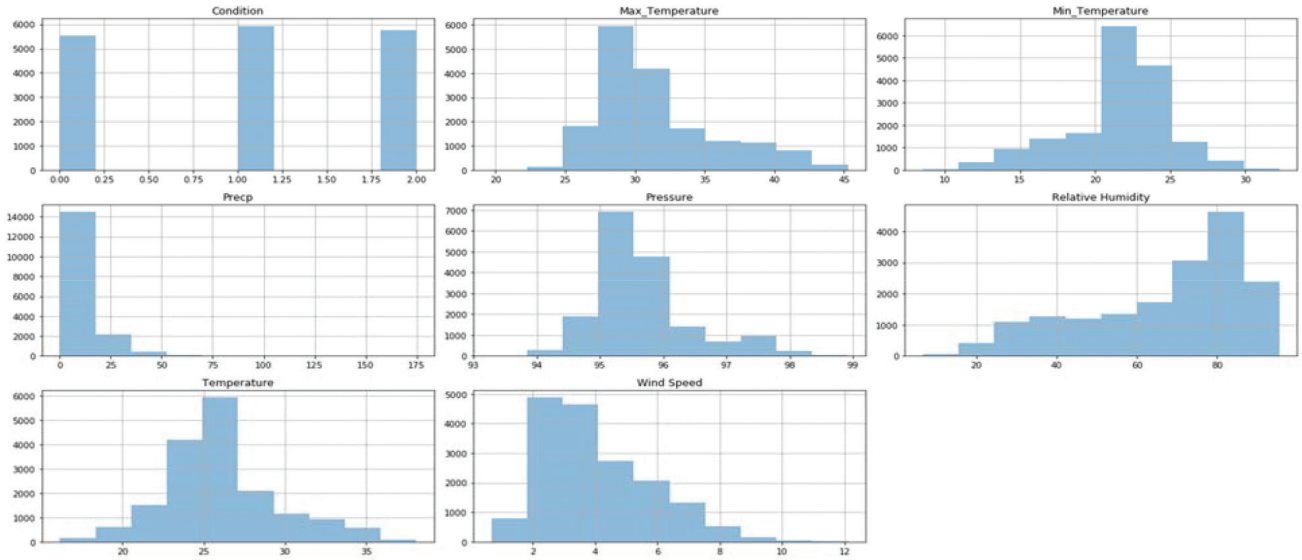

Figure 3: Considered attributes (parameters) in dataset.

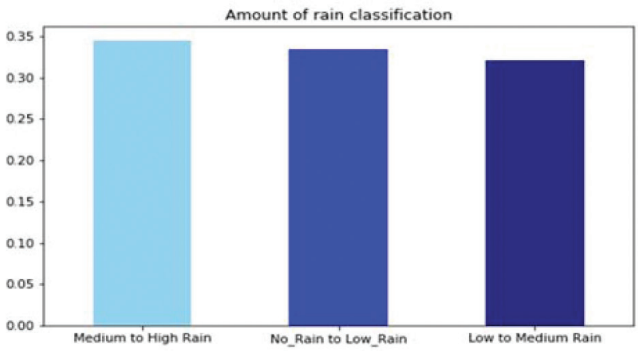

Figure 4: Amount of data in each category.

- Very High rain — when the precipitation rate is > 50 mm (2.0 in) per hour

As per the experiment, the dataset is divided into three categories, as described in Table 2

Data balancing is an essential step in data pre-processing. If the data is not balanced, the trained model will show predictions biased to the category with the highest data. Balanced data means there should not be much difference in the dataset's amount of data concerning each category. In this experiment, the data is collected and manually pre-processed to get balanced data in each category, as shown in Figure 4.

Table 2: Division of dataset based on the amount of precipitation

Classification	*Condition*	*Data in rows*
No rain to very Low Rain	Precipitation ≥ 0 mm and Precipitation < 0.1 mm	5751
Low to Medium Rain	Precipitation ≥ 0.1 mm and Precipitation < 10 mm	5531
Medium to High Rain	Precipitation ≥10 mm	5937

Data Pre-processing

Data pre-processing is one of the vital steps in any ML model as the training of the model depends upon the type of data, features, their correlation, and filling of missing values if any; scaling or normalization are some of the steps that make the model more powerful and robust (Sunitha et al., 2016).

Data mining is the process of analysing the data and extracting important information from a huge amount of data that helps in decision-making by mining the hidden predictive data from a large amount of data. It mines vast databases for hidden prognostic information. It's a promising new methodology with much data analysis and decision-making potential. Rainfall prediction is the state of the atmosphere that can be predicted by applying science and technology by collecting various weather parameters like temperature, relative humidity, dewpoint, atmospheric pressure, wind speed, etc.

Handling Missing values, if any, is very important. Sometimes some parameters were not recorded and were found missing in the dataset. Missing values lead to problems like the reduction of statistical power due to the absence of value. Sometimes missing data end up in biased estimation. The presence of missing values will drastically impact the performance of machine learning models, especially in deterministic models. Different ways of handling missing values are:

1. Discarding rows containing missing values will result in loss of information if missing values are in large numbers.
2. Missing values can be filled by calculating the domain's mean or median.
3. Various strategies like oversampling, under-sampling, or SMOTE are available to handle missing values.

Label Encoding is required to represent textual data in numerical form. The computer understands only binary data, so any textual data must be represented as numbers using label encoding. Label Encoding converts labels into numbers from 0 to no_of_classes minus 1.

One-Hot-Encoding further converts a numerical value into binary format (vector) where all the classes will be given 0 except the class having value one to avoid the priorities for taking the input correlation.

Feature Scaling: The most important feature is the normalisation or scaling of data for many machine learning algorithms. Feature scaling is used to handle the outliers and to keep the data in the proper range. Normalisation and standardisation are popular approaches for scaling numerical data before modeling.

Normalisation scales each input variable separately to the range 0-1, which is the most precise range for floating-point data. Standardisation shifts the distribution to have a mean of zero and a standard deviation of one by subtracting the mean (called centering) and dividing it by the standard deviation for each input variable.

Many more pre-processing techniques are available, but the steps mentioned above are most important and are efficiently used in the dataset chosen and applied wherever required.

Methods and Methodologies

Logistic Regression: Logistic regression is a mathematical formula-based process that predicts a binary outcome: either something happens, or nothing happens. Logistic regression should be avoided if the amount of data is less than the features as it may lead to overfitting (Moon et al., 2019). It is a viral ML algorithm used in many classification problems like cancer detection, detection of spam, etc., and it is best for figuring out how numerous independent variables influence a single result variable.

Decision Tree: A DT is a supervised learning algorithm that works on the dataset by breaking it into smaller subsets and incrementally develops an associated decision tree. A decision tree comprises many branches and nodes, and the leaf nodes (end nodes) store the final result and represent a decision. The root node is the topmost node in the decision tree and corresponds to the best predictor. A decision tree requires less data pre-processing and can handle both numerical and categorical data efficiently, and is simple to implement and visualise the data (Ramsundram et al., 2016). Decision trees can produce complicated trees that are difficult to generalise and unstable since slight changes

in the data might result in the generation of an entirely different tree.

Random Forest: RF algorithm is an ensemble technique with many decision trees (DT) grouped with training data, and then new data is fitted within one of the trees as a "random forest". Here many decision trees as weak learners come together to become strong learners as a random forest and give a robust model as output (Zainudin et al., 2016). The overfitting problem is very efficiently handled by a Random Forest classifier and found to be more accurate than the decision tree model in several cases but requires more execution time; thus, real-time prediction becomes slow and sometimes may face difficulty in implementing complex algorithms.

Naive Bayes: Naive Bayes calculates the likelihood of a data point falling into a specific category or not. It could be used in text analysis to classify words or phrases as belonging to a current "tag" (classification) or not. It's a variant of the Bayes theorem in which each feature is assumed to be independent. This technique takes less training time to estimate the relevant parameters and performs extremely fast but gives poor estimation compared to other classifiers and is known to be a bad estimator.

K-Nearest Neighbours: KNN is a pattern recognition technique that finds the k closest relatives in future cases using training datasets. When used for classification, K-NN calculates the data in the appropriate category of its nearest neighbour. Depending on the value of k, it would be placed in the class nearest to the k value. Classification of k is decided by the plurality of votes of its neighbours. Determining the k value makes the computational cost high as the distance needs to be calculated for each instance to all training samples. This algorithm was advantageous when dealing with massive training data and is robust and straightforward to implement.

Support Vector Machines: SVM uses algorithms to train and classify data within degrees of polarity, taking it to a degree beyond X/Y prediction. We find the ideal hyperplane that differentiates between the two classes, and the hyperplane is drawn with the help of a support vector. The margin of the classifier is maximised using these support vectors. This algorithm can be used both for classification and regression problems. It works well in high-dimensional areas and only utilizes a small number of training points in the decision function, making it memory-friendly. The algorithm does not provide probability estimates directly; instead, they are obtained through a time-consuming five-fold cross-validation procedure (Rajasekhar et al., 2014).

Ensemble Techniques: Ensemble techniques are used to boost the performance of ML models by combining the decisions from multiple models. Mainly used, three main ensemble techniques are:

- **Bagging:** The technique of fitting numerous decision trees to different samples of the same dataset and then averaging the results is known as Bagging. The best example of Bagging is Random Forest.
- **Boosting:** Boosting approaches create base estimators in sequential order, with each successive estimator learning from the errors of the one before it.
- **Stacking** is the process of fitting multiple models, referred to as base model (Level-0), to the same data and then using another model called a meta-model (Level-1) to learn how to integrate the predictions in the best way possible. The working of all main ensemble models is shown in Figure 5.

Basic Framework for Machine Learning Model

Data pre-processing is done by analysing the data collected from the Nasa power data portal for Hyderabad city, and all the collected weather parameters are observed. Manually the data is labeled based on the amount of precipitation as no rain to very low rain, low to medium rain, and medium to high rain. Figure 6 shows several steps to train the model, from collecting the dataset to cleaning it, feature extraction, and finally partitioning it into train and test sets. Next, several classification models are applied to the training set, and the performance is evaluated by computing the accuracy based on the test dataset predictions.

Platform and Libraries Used

Python is open source and a widely used general-purpose programming language with a wide range of applications and is accessible to everyone, easy to learn and understand, requires fewer lines of code with more considerable functionalities, and has significant developer communities. Python is excellent for data science, machine learning, and deep learning programming. Python has an excellent repository in terms of libraries**.** One of the greatest solutions for executing matrix operations computationally is the *NUMPY* library. Multi-dimensional arrays are supported. The *PANDAS* module is a Python open-source framework for creating data frames that are particularly useful for data organisation and is widely used to organize data systematically in data science,

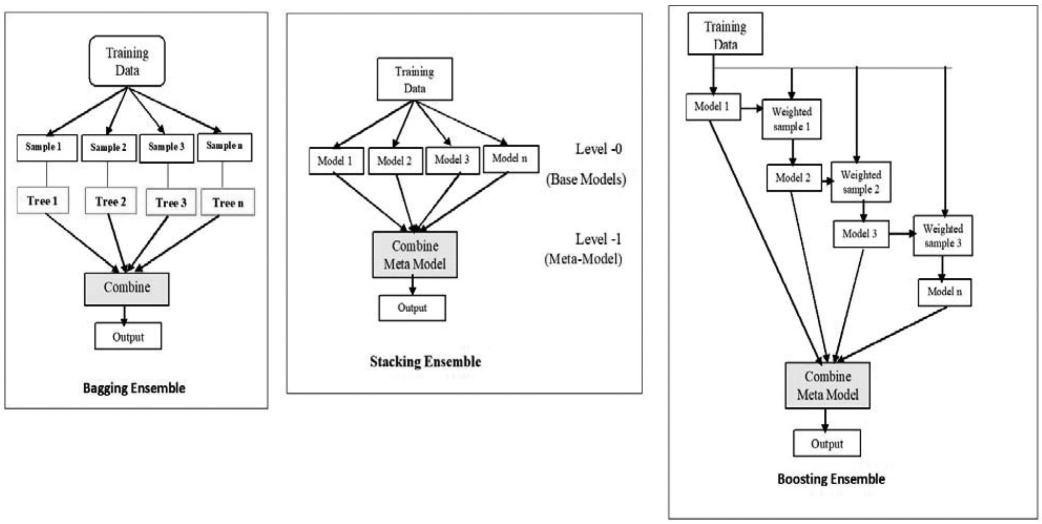

Figure 5: Three main ensemble models.

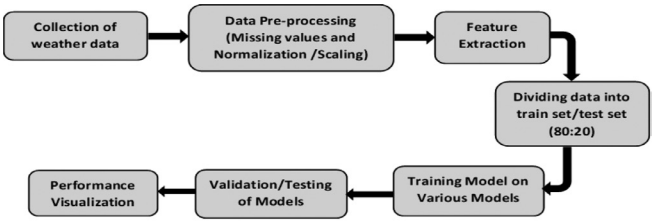

Figure 6: Basic structure of machine learning model.

machine learning, and deep learning. One of the best tools for visualizing data frames or any other type of data is the matplotlib module. In data science, Matplotlib is used to visualise data for exploratory data analysis. It helps in understanding the type of data we are dealing with and determining the following steps to take are both incredibly beneficial. SEABORN is a Python module for creating statistical visuals. It is mainly used to visualise the relationship between various variables in the dataset using its plotting function. One of the best tools for machine learning and predictive data analysis is the scikit-learn module, also known as sklearn. It has a lot of built-in algorithms, including logistic regression, support vector machines (SVMs), classification techniques like K-means clustering, and a lot more.

Results and Analysis

The experiment uses machine learning models, such as Logistic Regression, Decision Tree, Random Forest, K-nearest Neighbours, Support Vector Machine, and Naïve Bayes. Later ensemble techniques are also applied to train the model.

Metric of Evaluation: A very effective technique known as K-FOLD CROSS VALIDATION is used to evaluate ML models. The model is evaluated on a portion of the dataset that was not utilised for training. The entire dataset is partitioned into k-subsets, with one of the k subsets serving as a test/validating set and the remaining k-1 subsets serving as the training set. So it is very efficient and effective to estimate the model's performance on the unseen data as it is not used during the model's training. It results in less biased estimates as most of the data is used for training, and interchanging test sets increase its effectiveness. The preferable value of k is 5 or 10. The Accuracy obtained by various models using the K-fold cross-validation technique is shown in Table 3 and is visualised in Figure 7 using boxplots. In this experiment, k=5 is used.

Table 3: Accuracy obtained by various models

Model	*Accuracy obtained*
Naïve Bayes	0.665
Decision Tree	0.690
Logistic Regression	0.698
Support Vector Machine	0.709
K-Nearest Neighbour	0.721
Random Forest	0.758
Voting Classifier (Hybrid)	0.741
Stacking Ensemble	0.761

It was found that, when compared to other individual machine learning models, RF and KNN provided the best predictions. While logistic regression and SVM fared well in predicting values, DT and Nave Byes

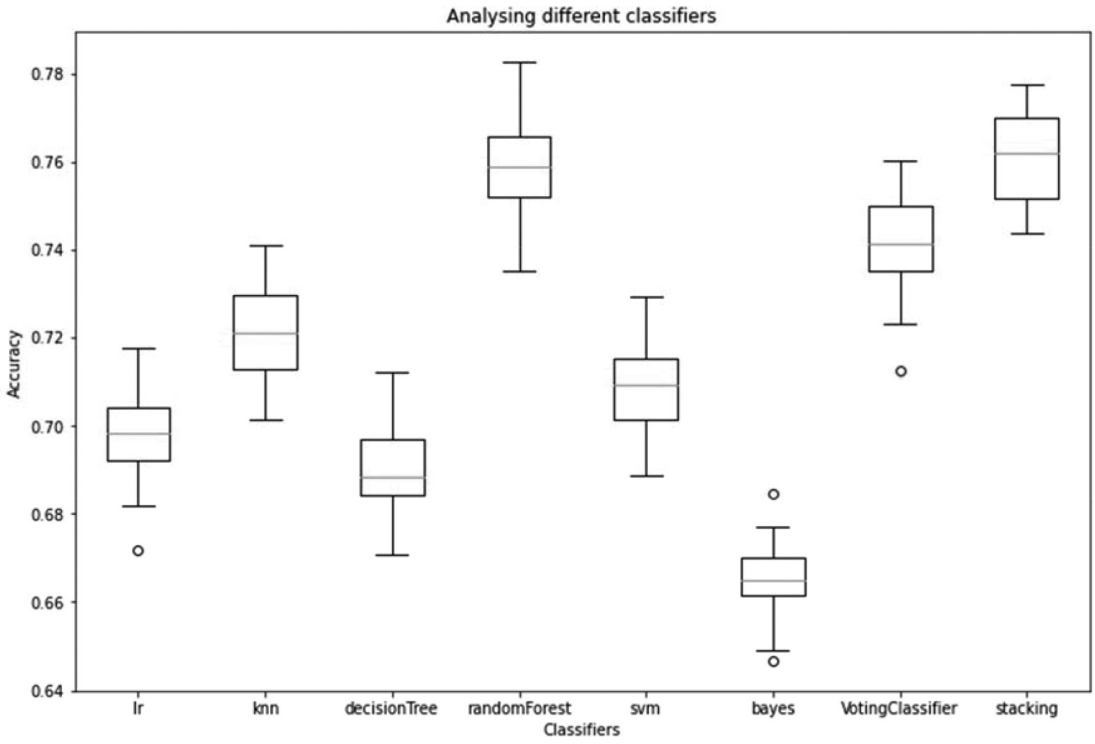

Figure 7: Accuracy obtained by individual classifiers and ensembles.

did poorly estimating predictions compared to other models. Predictions from stacking ensemble are better than voting, but the calculation time is three times longer than for individual machine learning models. In stacking, ensemble base models are six, as mentioned above, which are at level 0, and the meta-model at level -1 is taken as logistic regression in this experiment.

Conclusion and Future Scope

Thus, rainfall prediction is significant in many fields like agriculture, planning outdoor events, and transporting goods over long distances through various locations. The machine learning model helps to predict the intensity of rainfall by knowing the current atmosphere parameters at that location which are trained on the data accumulated over many years. In general, the designed model performance is excellent. It has the potential to predict the intensity of rainfall as no rain, low to medium rain, or medium to high rain based on the current atmospheric parameters at a particular location which can be used in various decision-making activities based on the status of rainfall and will be very much helpful to the people working in the sector that depends on rainfall like agriculture, tourism, transportation, etc.

The rainfall prediction can be enhanced in the future by classifying the precipitation into more categories to determine the precise rainfall by expanding the dataset for different locations. Additionally, the atmospheric data can be combined with images of the sky to identify the types of clouds present to make the rainfall predictions more promising as the clouds play an important role in the rainfall.

References

Abdullah, S. and Ismail, M., 2019. The weather and climate of tropical TasikKenyir, Terengganu. *In*: Greater Kenyir Landscapes (pp. 3-8). Springer, Cham. https://doi.org/10.1007/978-3-319-92264-5_1

Agovino, M., Casaccia, M., Ciommi, M., Ferrara, M. and Marchesano, K., 2019. Agriculture, climate change and sustainability: The case of EU-28. *Ecological Indicators*, **105:** 525-543. https://doi.org/10.1016/j.ecolind.2018.04.064

Appiah-Badu, N.K.A., Missah, Y.M., Amekudzi, L.K., Ussiph, N., Frimpong, T. and Ahene, E., 2021. Rainfall prediction using machine learning algorithms for the various ecological znes of Ghana. *IEEE Access*, **10:**pp.5069-5082. DOI: 10.1109/ACCESS.2021.3139312

Cahir, J.J., 2013, Weather Forecasting. Encyclopedia Britannica. Accessed on June 2013 (http://www.britannica.com/EBchecked/ topic/638321/weather-forecasting).

Cramer, S., Kampouridis, M., Freitas, A.A. and Alexandridis, A.K., 2017. An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *Expert Systems with Applications*, **85:** 169-181. https://doi.org/10.1016/j.eswa.2017.05.029

Didal, V.K., Brijbhooshan, A.T. and Choudhary, K., 2017. Weather forecasting in India: A review. *Int. J. Curr. Microbiol. App. Sci*., **6(11):** 577-590. doi: https://doi.org/10.20546/ijcmas.2017.611.070

Gutierrez-Lopez, A., Cruz-Paz, I. and Muñoz Mandujano, M., 2019. Algorithm to predict the rainfall starting point as a function of atmospheric pressure, humidity, and dewpoint. *Climate*, **7(11):** 131. https://doi.org/10.3390/cli7110131

Holley, D.M., Dorling, S.R., Steele, C.J. and Earl, N., 2014. A climatology of convective available potential energy in Great Britain. *International Journal of Climatology*, **34(14):** 3811-3824. https://doi.org/10.1002/joc.3976

Jaseena, K.U. and Kovoor, B.C., 2020. Deterministic weather forecasting models based on intelligent predictors: A survey. *Journal of King Saud University-Computer and Information Sciences*, **34(6B):** 3393-3412. https://doi.org/10.1016/j.jksuci.2020.09.009

Linacre, E. and Geerts, B., 2002. *Climates and Weather Explained*. Routledge.

Lekouch, I., Lekouch, K., Muselli, M., Mongruel, A., Kabbachi, B. and Beysens, D., 2012. Rooftop dew, fog and rain collection in southwest Morocco and predictive dew modeling using neural networks. *Journal of Hydrology*, **448:** 60-72. https://doi.org/10.1016/j.jhydrol.2012.04.004

Manandhar, S., Dev, S., Lee, Y.H., Meng, Y.S. and Winkler, S., 2019. A data-driven approach for accurate rainfall prediction. *IEEE Transactions on Geoscience and Remote Sensing*, **57(11):** 9323-9331. https://doi.org/10.1109/tgrs.2019.2926110

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M. and Nauss, T., 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, **101:** 1-9.

Misra, U., Deshamukhya, A., Sharma, S. and Pal, S., 2018. Simulation of daily rainfall from concurrent meteorological parameters over core monsoon region of India: A novel approach. *Advances in Meteorology*, **2018:** 3053640. https://doi.org/10.1155/2018/3053640

Moon, S.H., Kim, Y.H., Lee, Y.H. and Moon, B.R., 2019. Application of machine learning to an early warning system for very short-term heavy rainfall. *Journal of Hydrology*, **568:** 1042-1054. https://doi.org/10.1016/j.jhydrol.2018.11.060

Narvekar, M. and Fargose, P., 2015. Daily weather forecasting using artificial neural network. *International Journal of Computer Applications*, **121(22):** 9-13. https://doi.org/10.5120/21830-5088

Nastos, P.T., Moustris, K.P., Larissi, I.K. and Paliatsos, A.G., 2013. Rain intensity forecast using artificial neural networks in Athens, Greece. *Atmospheric Research*, **119:** 153-160. https://doi.org/10.1016/j.atmosres.2011.07.020

Pereira, G.F., Balasubramanian, G., Sabarathinam, C., Goswami, S. and Swaminathan, B., 2022. Long term microscale decadal analysis of coastal rainfall pattern: An indication of microclimatic variation in South India. *Journal of Climate Change*, **8(2):** 7-22.doi: 10.3233/JCC220010

Rajasekhar, N. and Rajinikanth, T.V., 2014. Weather analysis of Guntur district of Andhra region using hybrid SVM data mining techniques. *International Journal of Engineering and Advanced Technology (IJEAT)*, **3(4):** 133-136.

Ramsundram, N., Sathya, S. and Karthikeyan, S., 2016. Comparison of decision tree based rainfall prediction model with data driven model considering climatic variables. *Irrigation and Drainage Systems Engineering*, **5(3):** 1-5. https://doi.org/10.4172/2168-9768.1000175

Sunitha, L., Balraju, M., Sasikiran, J. and Kumar, B.A., 2016. Finding relation between parameters of weather data using linear regression method. *International Journal of Research in Engineering and Technology*, **5(5):** 90-93. https://doi.org/10.15623/ijret.2016.0517020

Tharun, V.P., Prakash, R. and Devi, S.R., 2018. April. Prediction of rainfall using data mining techniques. *In:* 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1507-1512). IEEE.https://doi.org/10.1109/icicct.2018.8473177

Zainudin, S., Jasim, D.S. and Bakar, A.A., 2016. Comparative analysis of data mining techniques for Malaysian rainfall prediction. *Int. J. Adv. Sci. Eng. Inf. Technol*, **6(6):** 1148-1153. https://doi.org/10.18517/ijaseit.6.6.1487